ISCR

Institute for Scientific Computing Research

University Collaborative Research Program Subcontract Research Summaries

CASC Center for Applied Scientific Computing

Application of Probe-Based Storage to High-Performance Computing

Zachary Peterson and Darrell Long

University of California, Santa Cruz

Summary:

Although there is continual improvement in speed and capacity of storage systems, traditional longitudinal magnetic recording is approaching a hard physical limit. A performance gap between the speed and capacity of RAM and disk is increasing at a rate of 50% a year. This performance gap is especially visible in high-performance workloads such as those employed in NNSA's Advanced Simulation and Computing Program (ASCI).

We in the Computer Systems Research Group (CSRG) at the University of California, Santa Cruz, are investigating two lines of new and innovative high-performance storage research. The first involves an exciting new storage technology based on MicroElectroMechanical Systems (MEMS), which promises a significant increase in performance, capacity, and reliability relative to modern storage devices. The second approaches the performance problems of current storage devices, applying novel ideas of grouping like data to improve read performance, while eliminating the amount of rewriting needed to be done to keep these data contiguous. Using sanitized workload traces from LLNL and support from the ISCR program, we have shown both lines of research successful in addressing the I/O problems posed by ASCI applications.

Recent advances in MEMS technology have given light to exciting new possibilities in storage technology. Using arrays of tiny read/write heads and electrostatic activated microactuators in conjunction with media that moves two dimensions, MEMS-based storage promises huge advantages over contemporary magnetic media. These advantages include: a higher bit density per square inch, a storage of less mass that requires less power to operate, an efficient device-based redundancy of data, and parallelism that increases the data access rate to up to 40 times that of current storage devices. The CSRG has led investigative research in these innovative devices. We have begun to examine the low-level characteristics of MEMS-based storage through the use of simulated device modeling, which enables us to better understand how this new technology can best serve a high-performance computing environment. The CSRG has three device models that address the performance memory gap in different ways: as a replacement for mass secondary storage devices, as a new layer of cache in the memory hierarchy, and as a component of a storage subsystem acting as a large read/write buffer for "hot" metadata and data. Each of these device applications addresses a different aspect of the memory performance gap, meaning all models are feasible in a single high-performance system, providing high throughput and reduced latency.

Beyond accurate and innovative device modeling, we continue our investigation by looking at how I/O scheduling affects device performance. We have been able to not only apply existing policies and algorithms, but also to develop new scheduling algorithms that take advantage of the unique characteristics of this implicitly parallel device. By varying sled movement and the way data is placed in the device, we can better serve I/O requests both satisfying fairness and performance requirements. Our results show that

Summary (continued):

Zachary Peterson and Darrell Long
University of California, Santa Cruz

algorithms developed here in the CSRG outperform almost all algorithms intended for standard magnetic media. It is clear that an application of existing technology to this device is not a sufficient or acceptable solution.

We have been able to make excellent progress on modeling, architectural alternatives, and I/O scheduling. However, there is still research to be done in these areas as well as in aspects of data layout to fully take advantage of the exciting possibilities offered by MEMS-based storage.

Email: zachary@cse.ucsc.edu

Personal web page: http://people.ucsc.edu/~zacharyp/

Institution web page: http://www.ucsc.edu/

Mining Large Image Datasets

Jelena Tesic and Bangalore Manjunath

University of California, Santa Barbara

Summary:

Mining large image datasets places a number of challenging requirements on the analysis framework. Some initial success has been achieved with systems that represent images as an organized collection of summarized information obtained from the feature descriptors and spatial constraints. However, the high dimensionality of the feature spaces and the size of the image datasets make meaningful summarization a challenging problem. A visual thesaurus based on low-level image descriptors provides a scalable conceptual framework for analyzing perceptual events. The heart of this method is a learning system that gathers information by interacting with database users.

Our main objective is to find clusters that represent similar feature points located in a small subset of a feature space. High-dimensional spaces represent challenges for clustering, due to the sparseness of the space. However, clusters may be formed from a couple of visually different elements that inhabit a large part of a high-dimensional space. Co-occurrence of clusters in an image helps us distinguish visually meaningful representatives. We are currently conducting experiments on texture feature sets to determine the dependency of the clusters in the texture feature space on feature vectors and spatial image layout.

The objective of the visual thesaurus is to classify the image regions into perceptually similar categories. Spatial Event Cubes (SECs) are used to represent and analyze the spatial relationships. SECs are computed with respect to particular spatial relationships. Detailed analysis shows that SECs can be used for visualization, discovery of latent spatial configurations, and for constructing indices for efficient and meaningful data access.

Email: jelena@ece.ucsb.edu

Institution website: http://www.ucsb.edu/

Personal website: http://vision.ece.ucsb.edu/~jelena/

Sensitivity and Uncertainty Analysis for Large-Scale Differential-Algebraic Systems

Yang Cao and Linda Petzold

University of California, Santa Barbara

Summary:

Sensitivity analysis generates essential information for design optimization, parameter estimation, optimal control, data assimilation, process sensitivity, and experimental design. In our earlier work, the DASPK3.0 software package was developed for forward sensitivity analysis of differential-algebraic (DAE) systems. Some problems require the sensitivities with respect to a large number of parameters. For these problems, particularly if the number of state variables is also large, the forward sensitivity approach is intractable. We have developed an efficient algorithm for sensitivity computation of large-scale differential-algebraic systems based on the adjoint method. The new algorithm is more efficient than forward sensitivity analysis when there are a large number of parameters. We have analyzed the issues of stability of the adjoint problem and its numerical solution, the determination of consistent boundary conditions for the adjoint system, and the conditions under which the sensitivity analysis problem is well posed. Software called DASPKADJOINT based on (an extended version of) DASPK3.0 and our new adjoint method was developed.

We are finding that the adjoint method is a very powerful tool for the estimation of computational and modeling errors in general, and we have begun work on applying it to the estimation of errors of reduced and/or simplified models. In a surprising twist, we have discovered an adjoint-based condition estimator for a simpler class of problems: solution of linear systems of equations. The new condition estimator appears to be more accurate and efficient than current techniques, and is also adaptable to a greater variety of needs. We have also begun work on both forward and adjoint sensitivity analysis for PDEs. The adjoint problem for PDEs has many challenges, due to the difficulties in finding the adjoint for general problems with different boundary conditions, and also when using nonlinear and adaptive discretization schemes.

Email: petzold@engineering.ucsb.edu

Institutional website: http://www.ucsb.edu/

Personal website: http://www.engineering.ucsb.edu/~cse/

Simulation of Compressible Turbulent Flows With Reaction

David Lopez, Carlos Pantano, and Sutanu Sarkar

University of California, San Diego

Summary:

Two outstanding problems in the area of compressible reactive flows are being investigated. The first problem concerns large eddy simulation where the strongly nonlinear dependence of the reaction rate term on temperature leads to subgrid contribution in the resolved-scale equations, which must be modeled. We have developed a subgrid model by using the information available in the resolved scales along with additional physical input, namely, a model spectrum for the unresolved scales. Promising results have been obtained when evaluating this subgrid reaction rate model against a direct numerical simulation of a shear layer. The second problem occurs when a burn initiated in a NIF capsule encounters inhomogeneities of mixture fraction due to the introduction of inert shell material into DT mix by Rayleigh–Taylor or Richtmeyer–Meshkov instabilities. Preliminary results have been obtained in a simple model problem to identify the important parameters that control the modified burn propagation.

Email: sarkar@mechanics.ucsd.edu

Personal web page: http://www-mae.ucsd.edu/RESEARCH/SARKAR/sarkar.html

Institution web page: http://www.ucsd.edu/

Probabilistic Clustering of Dynamic Trajectories for Scientific Data Mining

Scott Gaffney and Padhraic Smyth

University of California, Irvine

Summary:

Data-driven exploration of massive spatio-temporal data sets is an area where there is particular need of new data analysis and data mining techniques. Analysis of spatio-temporal data is inherently challenging, yet most current research in data mining is focused on algorithms based on more traditional feature-vector data representations. The goal of this research is to develop a flexible and robust framework (as well as algorithms and software tools) for tracking and clustering time-trajectories of coherent structures in spatio-temporal grid data. The benefit of such an approach is to begin to provide a basic set of data analysis tools for exploration and modeling of dynamic objects, in a manner analogous to the much more widely available techniques for clustering of multivariate vector data (e.g., k-means, Gaussian mixtures, hierarchical clustering, etc).

To date in this work we have investigated a general probabilistic approach to clustering of sets of trajectories. We assume that the set of trajectories was generated by a mixture of K component trajectory models, where each component model describes a particular type of trajectory. We have implemented in MATLAB a flexible EM-based clustering algorithm based on these ideas. The algorithm takes as input a set of trajectories, where each is described as a multivariate vector of measurements over time, e.g., (x,y) position and/or various object features measured at each time *t*. The trajectories can be of different time-durations and can have measurements at different times. Different forms of component trajectory model for each cluster can be selected by the user. We have currently implemented the following cluster models: linear regression, polynomial regression, and non-parametric kernel regression. The algorithm is also supplied with a value for *K*, the desired number of clusters. We have also implemented a cross-validation algorithm that selects the best value for *K*, the number of clusters, based on cross-validated predictive probability scores. We have tested the models and the general algorithm on a number of simulated data sets and verified its ability to recover the underlying *K* data-generating trajectory component models from a set of *N* trajectories without labels.

The primary application of trajectory clustering that we have investigated up to this point is the clustering of extra-tropical cyclone (ETC) tracks from meteorological data over the earth's northern hemisphere. ETCs are significant for a number of reasons. For example, they are responsible for severe and highly damaging weather over North America and western Europe. In addition, it is not yet well-understood how ETC patterns are correlated with long-term climatic phenomena and what the implications of global warming might be on ETC frequency, intensity, and spatial distribution. Thus, quantification of the spatial and temporal patterns of ETCs is well motivated.

We have analyzed 15 winters of the mean sea-level pressure field for the CCM3 AMIP II data set, at 6-hourly intervals, on a 2.5-degree square grid over the North Atlantic. Using this data we have developed an algorithm (and implemented it in MATLAB) to locate and track candidate cyclone centers in each 2D frame of the data. The resulting cyclone trajectories were then used as input to our MATLAB model-based clustering algorithm. Analysis of the results suggest that a *K*=3 model with linear regression components provides both a good fit to the data as well as being quite interpretable from an atmospheric science

Summary (Continued):

Scott Gaffney and Padhraic Smyth

University of California, Irvine

viewpoint. We are currently in the process of analyzing the clustering results with atmospheric science collaborators at UCLA (Andy Robertson and Michael Ghil). From a scientific viewpoint there are a number of open questions in terms of interpretation of the cluster models: How should the results be visualized? Can we analyze intensity profiles as well positional information? How predictable are the cyclone trajectories? Are the clusters correlated with other, more global atmospheric phenomena such as the North Atlantic Oscillation index? We are currently in the process of investigating these questions.

On a more general note, we are also investigating how our methodology can be generalized to other types of trajectory data analysis problems. For example, the spatial information (in the form of morphological features) is of interest in many applications and a natural question is how to incorporate this in the clustering. Another question of interest is the sensitivity of our methodology to the particular component models being used: if the trajectories are not truly linear, but linear component models are used, how does this affect the quality of the results? We are investigating the use of statistical random effects models as a methodology to generalize in the direction of a more flexible model.

Email: smyth@ics.uci.edu

Institutional website: http://www.uci.edu/

Personal website: http://www1.ics.uci.edu/~smyth/